SemGen—Towards a Semantic Data Generator for Benchmarking Duplicate Detectors*

Norbert Baumgartner[1], Wolfgang Gottesheim[2], Stefan Mitsch[2],
Werner Retschitzegger[2], and Wieland Schwinger[2]

[1] team Communication Technology Mgt. Ltd., Goethegasse 3, 1010 Vienna, Austria
[2] Johannes Kepler University Linz, Altenbergerstr. 69, 4040 Linz, Austria

Abstract. Benchmarking the quality of duplicate detection methods requires comprehensive knowledge on duplicate pairs in addition to sufficient size and variability of test data sets. While extending real-world data sets with artificially created data is promising, current approaches to such *synthetic data generation*, however, work solely on a quantitative level, which entails that duplicate semantics are only implicitly represented, leading to only insufficiently configurable variability.

In this paper we propose SemGen, a semantics-driven approach to synthetic data generation. SemGen first diversifies real-world objects on a *qualitative level*, before in a second step quantitative values are generated. To demonstrate the applicability of SemGen, we propose how to define duplicate semantics for the domain of road traffic management. A discussion of lessons learned concludes the paper.

1 Introduction

Duplicate detection is an elementary part in data cleansing processes and addresses the identification of multiple different representations of one and the same real-world object within a data set [16]. Such cleansing processes are vital components in information systems that integrate multiple data sources, as it is the case in systems that support situation awareness. We are currently developing a framework for realizing ontology-driven situation awareness techniques [2], including duplicate detection techniques [3], in the sample domain of road traffic management. Real-world objects are described by *object representations characterized by attributes* that specify their spatial and temporal extent, for example in the form of a region on a highway defined by a start and end point. From such attributes, qualitative relations between objects can be derived that characterize various aspects of objects. For example, from a spatial perspective such aspects could be size, distance, or mereotopology of objects. In situation awareness, spatio-temporal data on objects is incrementally reported in streams, describing real-world evolution courses. Within these data, duplicates may occur in multiple forms (see [15] for a taxonomy on the subject). Most relevant

* This work has been funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant FIT-IT 819577.

for our domain are those arising from *identical attribute values* (e. g., two representations of the same traffic jam with the same regions), *contradictory values* (e. g., two representations of the same traffic jam differ in terms of spatial extent, which may be caused by measuring or entry errors), or *missing values* (e. g., only the start values of the region of a traffic jam are given). Duplicate detection is therefore typically performed by computing similarity measures for pairs of representations on a per-attribute basis, which are aggregated into an overall duplicate decision [16].

A semantics-driven approach to synthetic data set generation. We have defined the following three requirements for a test data generator that provides synthetic data sets for testing duplicate detection methods: (i) *Variability* within the generated data set has to be configurable with regard to multiple aspects to support testing effectiveness. This entails providing accurate numbers on generated duplicates to allow the computation of measures such as precision and recall. (ii) *Distributions* within an aspect in the generated data sets have to be configurable, enabling testing duplicate detection methods for robustness. (iii) *Different quantitative representations* should be realizable so that multiple duplicate detection methods can be tested. For instance, in the domain of road traffic management duplicate detection methods might be required to interpret regions with their spatial extent specified either in kilometers or with a distance measures basing on nodes in a graph describing highway exits. Therefore, quantitative representation for both cases have to be generated.

In this paper we propose SemGen, a semantics-driven approach to synthetic data generation. It is based on a qualitative definition of duplicate semantics and requires a set of data with pairs of objects marked as duplicates of each other—in the following called *labelled duplicates*—and non-duplicates, which are both first diversified on a qualitative level according to duplicate semantics of a domain, before in a second step quantitative values are generated, thereby enabling the creation of data sets with high variability and in different sizes.

Structure of the paper. In Section 2 we detail on qualitative descriptions of duplicate semantics, before we describe our approach in Section 3. Section 4 discusses relevant related work on synthetic test data generation, and finally Section 5 concludes the paper with a discussion of its findings and an outlook on further work.

2 Qualitative Description of Duplicate Semantics

Describing duplicate semantics using spatio-temporal relations on a qualitative level has been proposed as a basis for duplicate detection in our previous work [3]. In the following, we provide an overview on these qualitative descriptions and show how to use them for controlling variability and distribution in a generated data set.

Qualitative relations between two objects are expressed by employing *relation calculi*, each of them focusing on a certain spatio-temporal aspect, such as mereotopology [17], orientation [9], or temporality [1]. These calculi are often

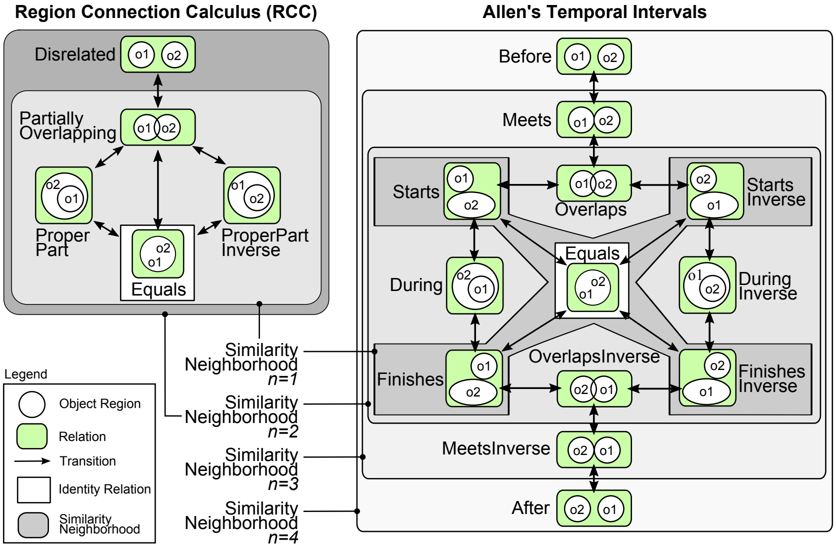

Fig. 1: Conceptual neighborhood graphs of RCC and Allen's Temporal Intervals.

formalized by means of *Conceptual Neighborhood Graphs* (CNGs, [10]), which originate in the field of spatio-temporal reasoning. Sample graphs for the Region Connection Calculus (RCC, [17]) and Allen's Temporal Intervals algebra [1] are shown in Fig. 1. In addition, CNGs define similarity between relations since, according to [11], relations are "*conceptual neighbors* if a direct transition from one relation to the other can occur upon an arbitrarily small change in the referenced domain" (e. g., *ProperPart* and *PartiallyOverlapping* are more similar to *Equals* than *Disrelated*). In each such relation calculus, one can define an *identity relation* [3], which states that two objects being in such a relation are most similar according to the particular calculus' aspect of the world (e. g., *rcc*:*Equals* is the identity relation of RCC, *allen*:*Equals* the one of Allen's Temporal Intervals). Qualitative relations between objects can be automatically derived from their quantitative attributes using rule-based relation interpretations [2] (e. g., two traffic jams are *PartiallyOverlapping* if their spatial regions overlap). We exploit these relations for describing in which aspects an object and its duplicate should be alike or different. While a number of holding identity relations shows that two objects are *duplicates from identical attribute values* with regard to these aspects, duplicates arising from *contradictory values*, i. e., values describing the same real world object in different ways, can be created by performing *qualitative diversification*. For example, if two objects are in a relation *allen*:*Equals*, their lifespans are the same, i. e., they "exist" at the same time. Still, they may differ, for instance, in a mereotopological aspect, described by the relation *rcc*:*PartiallyOverlapping* holding between them. Note that *duplicates arising from missing values* are a special case not reflected on the level of qualitative relations, because, as at this abstraction level we do not know which concrete attributes contribute to a relation, no statement can be made on missing attributes.

For describing duplicate semantics with this definition of CNGs and identity relations, we introduce the concept of *similarity neighborhoods*. A similarity neighborhood is defined by the set of relations reachable within n hops from the calculus' identity relation. Let us denote an instance of a particular object type O_i as a reference object $o_r^{O_i}$, and the similarity neighborhood around $o_r^{O_i}$ as $N_{calculus}(o_r^{O_i}, n)$. Synthetic objects with relations which are part of the similarity neighborhood (i. e., within n hops from the identity relation) are regarded as duplicates to the reference object, whereas objects outside the neighborhood are not labelled as duplicates. In Fig. 1, the similarity neighborhoods for RCC and Allen's Temporal Intervals algebra are shown.

By restricting n for each relation calculus to a particular value, we are able to steer qualitative diversification on a per-calculus basis. In addition, one could define n over multiple relation calculi, with the similarity of the synthetic object being defined by different aspects. For this, a generalization of relation neighborhood from one relation calculus to multiple calculi is necessary, which, however, has already been shown to be straightforward [7] by counting relations in the involved calculi. Finding an appropriate value for n is challenging and requires profound knowledge on the domain's properties. From our experience with duplicate detection in road traffic management we argue that, if using the CNGs of RCC and Allen's temporal intervals as shown above, $n = 2$ is a value yielding reasonable results. Nevertheless, it is desirable to include a larger number of different calculi, which in turn requires an adapted value for n. Using a total of $n = 2$ hops in the CNGs of RCC and Allen's temporal intervals, three different neighborhoods are reachable:

$N_{rcc}(o_r^{O_i}, 2)$**:** $\{rcc{:}Disrelated\}$
$N_{allen}(o_r^{O_i}, 2)$**:** $\{allen{:}Overlaps_{Inverse}, allen{:}During_{Inverse}\}$
$N_{rcc \wedge allen}(o_r^{O_i}, 2)$**:** $\{rcc{:}ProperPart, rcc{:}ProperPart_{Inverse}, rcc{:}PartiallyOverlapping\} \times \{allen{:}Starts_{Inverse}, allen{:}Finishes_{Inverse}\}$

Table 1 shows sample duplicates with relations holding between them and the similarity neighborhood they belong to.

To correlate quantitative values with their qualitative representations, we introduced rule-based relation interpretations that derive relations from object attribute values [2]. As a prerequisite, these relation interpretations assume that attribute values adhere to particular value ranges. These interpretations are domain-dependent, since the definition of such value ranges differs between domains. Again using road traffic management as a demonstration domain and representing mereotopological relations in RCC, let us demonstrate this concept. For deriving such mereotopological relations, using a strictly monotonic, linear space (i. e., road traffic objects, such as traffic jams or roadworks, that occupy a region on a highway) as value range, we can define regions as intervals, whereas given, for example, objects anchored in Euclidian space, we can define a region as a center point with a radius. We can now define the interpretations of relations in RCC ($rcc = \{Disrelated, PartiallyOverlapping, ProperPart, Proper$-

Table 1: Sample duplicates with their respective relations.

Qualitative Relations	**Quantitative Values**				
	ID	**Location**		**Time**	
		begin	end	begin	end
$rcc{:}Equal \wedge allen{:}Equals$: TJ_1' located in $N_{rcc \wedge allen}(TJ_1, 0) \rightarrow$ Duplicates (identical values)	TJ_1	km 6.5	km 8.0	2010-12-01 08:00	2010-12-01 09:00
	TJ_1'	km 6.5	km 8.0	2010-12-01 08:00	2010-12-01 09:00
$rcc{:}PartiallyOverlapping \wedge allen{:}Finishes_{Inverse}$: TJ_2' located in $N_{rcc \wedge allen}(TJ_2, 2) \rightarrow$ Duplicates (contradictory values)	TJ_2	km 7.5	km 11.0	2010-12-01 08:40	2010-12-01 09:00
	TJ_2'	km 8.0	km 13.5	2010-12-01 08:20	2010-12-01 09:00

(Legend: TJ = Traffic Jam)

$PartInverse, Equals\}$) as functions $f_{rcc} : \mathbb{R} \times \mathbb{R} \rightarrow rcc$ mapping object intervals to particular relations (e. g., $PartiallyOverlapping$ may be defined as $o1.start < o2.start \wedge o1.end > o2.start \wedge o1.end < o2.end$, as TJ_2 in Table 1 shows). For the purpose of data generation, we use the inverse of these functions, thereby mapping a qualitative relation onto a given value range. For example, for the above specified function f_{rcc} we can use its inverse $f_{rcc}^{-1} : rcc \rightarrow \mathbb{R} \times \mathbb{R}$ to map relations between two objects onto the underlying value range. The generation of duplicates arising from missing values can be performed here by providing an inverse function that either maps onto the value range or generates an empty result, such as $f_{rcc}^{-1} : rcc \rightarrow \mathbb{R} \vee \emptyset \times \mathbb{R} \vee \emptyset$.

Having laid the foundation for using qualitative data to describe how objects are related, the next section presents our approach to synthetic data generation exploiting the semantics of these relations.

3 Approach

SemGen, our semantics-driven approach to synthetic data generation creates duplicates by taking existing real-world duplicates as the basis for creating additional duplicates that closely resemble real-world characteristics. We control variability in the synthetic data set by using qualitative descriptions as outlined in the previous section. We propose a four-step process as depicted in Fig. 2:

1. *Relation derivation between labelled duplicate pairs* as starting points for the subsequent diversification steps,
2. *Qualitative diversification* to change these relations along the configured aspects,
3. *Quantitative diversification* to map the meaning of each such relation onto the attributes it is derived from, thereby finally characterizing synthetic objects in detail with sample attribute values derived from the attribute values of the labelled duplicate pair, and finally
4. *Export of generated synthetic objects* to an output format suiting the duplicate detection method to be evaluated, such as relational data or an ontology.

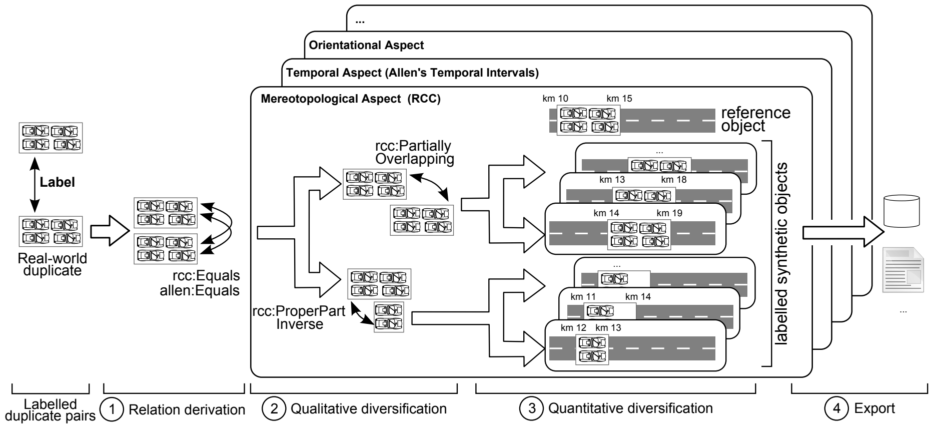

Fig. 2: Overview of synthetic duplicate generation.

Input for this process is a reference data set with *pairs of objects labelled as duplicates* as well as *distinct objects.* In road traffic management, such a data set contains information on objects in terms of quantitative attributes, such as spatial extent in kilometers or lifespan given as an interval of timestamps, and in terms of meta-information, such as the data source an objects originates from or an object's type. Common object types are, for example, traffic jams, road works, or lane closures.

Configuration mechanisms allow SemGen's user to control various aspects of the data generation process:

- the *size* of the generated data set,
- the kinds of *relation calculi* that are included in the process,
- the *distribution of relations* within a calculus,
- the *distribution of duplicates* in the whole data set,
- the *distribution of object types* in the resulting data set,
- and the *ratio of duplicates to non-duplicates* per object type.

Given our three causes for duplicates presented above—identical, contradictory, and missing information—it is obvious that the generation of identical duplicates is trivial, since the second step of this process can be omitted, therefore no values other than the original ones can be generated in step 3. Therefore, we focus on the generation of duplicates arising from contradictory and missing information.

(1) Relation derivation. As a prerequisite for the later process steps the relations between objects in a reference data set are derived. These relations serve as starting point for the subsequent qualitative diversification step. Currently, relation calculi relevant to the generation of spatio-temporal data are supported (RCC, spatial distance and size calculi [2], and Allen's Temporal Intervals).

(2) Qualitative diversification with CNGs. On the qualitative level, we employ the conceptual neighborhood of relations and identity relations introduced in Section 2 to define the concept of *similarity neighborhood* for steering qualitative diversification. We can now formulate an algorithm for choosing the relations (of particular aspects) which must hold between a given labelled reference object and its generated synthetic duplicate object, thereby respecting a similarity neighborhood constraint n for each relation calculus. Listing 1.1 shows a pseudocode representation of this algorithm[3]. In principle, the algorithm iterates over the relevant calculi and changes the relations between labelled reference objects and their synthetic duplicates in a random fashion, but constrained by the configured relation distribution. In order to determine the label of the synthetic object, i. e., whether the synthetic object is still a duplicate to its reference object, the algorithm checks whether or not the relation is within the neighborhood of the labelled reference object. If so, the synthetic object is assigned the same label as the reference object, otherwise it is not labelled. During this diversification step, the generation of relations that describe a single synthetic object in a conflicting and, hence, impossible way has to be avoided. An example for such a conflicting configuration occurs, if the *regions of two objects are equal*, while they are at the same time of *different size*.

Listing 1.1: Algorithm to find relations between a reference object and a synthetic object.

```
function select_relations(
  in configuration,
  in ref_object,
  in neighborhood_radii<calculus,n>,
  out relations_to_synth,
  out synth_labels)
  var neighborhood: set<relations>;
      rel: relation;
      relations_to_synth: set<relations>;
      synthetic_labels: set<label>;
  for each (calculus in neighborhood_radii.keys)
    neighbor_relations = N(object, neighborhood_radii[calculus]);
    repeat rel = random_select(neighbor_relations);
    until configuration.is_relation_acceptable(rel);
    if N.contains(rel) then
      // relation is in similarity neighborhood
      synth_label.add(get_label(ref_object));
    end if
    relations_to_synth.add(rel);
  end
end
```

(3) Quantitative diversification. In quantitative diversification, concrete values for object attributes need to be correlated with qualitative relations. For creating sample attribute values, we use the inverse functions to our relation derivation rules as introduced above. Concrete values for a sample region for a synthetic

[3] Note that, in this paper, the focus is put on showing its functional principle, thus ignoring possible performance improvements.

Table 2: Exemplary qualitative and quantitative diversification.

(a) Qualitative diversification.

			RCC			
			PO	EQ	PP	PP_i
		n	0	1	1	1
Allen's Temporal Intervals	$Finishes_{Inverse}$	0	$\circledast$			
	$During_{Inverse}$	1				
	$Overlaps_{Inverse}$	1				$\odot_{1..n}$
	$Equals$	1				
	$Starts$	1				
	$Finishes$	2		$\ominus_{1..n}$		
	$Starts_{Inverse}$	2				
	$Meets_{Inverse}$	2			$\circledcirc_{1..n}$	

(b) Quantitative diversification.

	ID	Location		Time	
		begin	end	begin	end
$\odot_1$	TJ_1	km 6.5	km 8.0	08:00	08:50
	TJ_2	km 7.0	km 7.5	08:40	09:00
$\odot_2$	TJ_1	km 6.5	km 8.0	08:00	08:35
	TJ_2	km 6.9	km 7.8	08:22	09:00
$\ominus_1$	TJ_1	km 6.5	km 8.0	08:30	09:00
	TJ_2	km 6.5	km 8.0	08:35	09:00
$\ominus_2$	TJ_1	km 6.5	km 8.0	08:47	09:00
	TJ_2	km 6.5	km 8.0	08:20	09:00
$\circledcirc_1$	TJ_1	km 8.0	km 10.0	08:30	09:00
	TJ_2	km 7.5	km 11.0	08:00	08:30
$\circledcirc_2$	TJ_1	km 8.1	km 9.9	08:45	09:00
	TJ_2	km 7.5	km 11.0	08:40	08:45

object are chosen randomly from applying the mapped interval to the region of the labelled reference object (e. g., a synthetic duplicate that is *ProperPart* of a labelled reference object has randomly chosen interval boundaries that lie within the boundaries of this labelled reference object). Since the relation calculi used for qualitative diversification are designed for reuse, interdependencies between them are not explicitly modeled. For example, the relation *rcc*:*Disrelated* does not specify in which order and at which distance objects are placed on the highway, leaving many options for quantitative diversification in a strictly monotonic, ordinal value space representing regions on a highway. In case several relation calculi, which describe the same real-world aspect, steer qualitative diversification, interdependencies between them put constraints on quantitative diversification. For instance, consider *rcc*:*Disrelated* and *spatdist*:*VeryClose* as a result of qualitative diversification. Then, sample attribute values created during quantitative diversification must satisfy both relation interpretations. To generate duplicates arising from missing values, random null values replacing sample attribute values can be generated during quantitative diversification to better mimic real-world data.

Example. To further illustrate the process described above, we will use two exemplary traffic objects TJ_1 and TJ_2 as shown in Table 1. As a minimal sample configuration based on our experience from road traffic management systems, we choose to use RCC as well as Allen's Temporal Intervals as relation calculi and configure a similarity neighborhood constraint of $n_{RCC} = 1$ and $n_{Allen} = 2$. Table 2a shows the resulting similarity neighborhood as a matrix.

In step (1), we derive relations of the configured calculi for our reference data set consisting, in this case, of two objects TJ_1 and TJ_2, which results in the relations $\{rcc{:}PartiallyOverlapping \wedge allen{:}Finishes_{Inverse}\}$ holding between TJ_1 and TJ_2 (denoted in Table 2a as $\circledast$). This means that their spatial regions overlap and, while the lifespan of TJ_2 begins after the one of TJ_1, both end at the same time. In step (2), these currently holding relations are diversified within the configured relation neighborhoods. This results in 31 possible additional configurations. Finally, in step (3), quantitative representations for these relations are generated based on the original attribute values, with some examples (denoted in Tab. 2a as $\odot$, $\ominus$, $\circledcirc$) shown in Table 2b. Note that attribute values affected in this process are highlighted.

4 Related Work

Automated generation of test data sets is an approach followed in a variety of fields. In the following, we will present domains where data generation approaches are used in order to show commonalities and differences to using data generators as a prerequisite for evaluating duplicate detection methods.

For database systems, Weis [18] distinguishes between data generators that facilitate tasks such as evaluating duplicate detection methods, and those that support the task of testing and improving the performance of a database system. We first cover closely related work from generators that belong to the first category, before we continue with more widely related approaches that fall into the second one. Among those, judging from literature the most well known for generating duplicates in relational data is *DBGen*, also known as *UIS Database Generator*[4], which manipulates records consisting of personal information such as name, address, and social security number by introducing typographical errors or completely changing them in a random fashion [12]. This approach has been refined in [4] to overcome some original limitations, such as poor variability in the set of possible values. Since both approaches are using implicit semantics for domains relying on string-based information, they do not allow to configure variability with regard to multiple aspects, and also lack support for multiple quantitative representations. thus suffering from the limitations described such as the lack of a semantically rich configuration mechanism allowing in-depth control of the generation process. Another approach from the first category is proposed in [8], where synthetic test data also containing duplicates is used to test applications using a relational database. Their goal was to create data sets that allow to verify the correct function of applications that access the database, and to that end, a comprehensive data set covering all relevant cases is required, which also includes the correct handling of duplicates. However, only identical duplicates are regarded, and furthermore configuration mechanisms as proposed here are missing. Thus, they are unable to configure variability with regard to multiple aspects, and do not support more than one quantitative representation.

Other data generators in the database field support the syntactical task of performance improvements by providing a large data set with known statistical properties in an efficient and reproducible way. In the last years, numerous approaches have been presented for generating data, such as [13], which provides efficient generation of large data sets in parallel and is flexibly configurable, or [6], which provides a “Data Generation Language” for specifying the generated data, or [14], using a graph model to control the generation process. But

[4] http://www.cs.utexas.edu/users/ml/riddle/data.html

their goal is to efficiently generate large data sets for performance testing which, therefore, have to be consistent and must not contain duplicates. Thus, using these generators for generating test data sets for duplicate detection methods is not feasible. In the area of spatio-temporal databases, frameworks to generate data on moving objects in a quantitative manner have been proposed [5], [**?**], which focus on generating data to represent the evolution of objects in terms of motion. While they are operating in a similar domain, again the focus is on the generation of consistent data, not of duplicates with known properties.

In summary, although synthetic data generation is an issue in many domains, qualitative approaches have not yet been the focus. Besides, many of these quantitative approaches are heavily domain-specific, limiting their applicability outside their original domain. To date, no data generator for the evaluation of duplicate detection methods in spatio-temporal data has been proposed.

5 Discussion and Further Work

In this section, we discuss several lessons learned during the ongoing implementation of the presented approach, which at the same time represent the directions followed by our further work.

Duplicate variability in real-world data configures qualitative diversification. By deriving relations between labelled duplicates in a small set of real-world data, distribution characteristics per relation calculus can be controlled on a qualitative level (e. g., in RCC, most duplicates may be `rcc:PartiallyOverlapping`, a smaller portion might be `rcc:Equals`, and some `rcc:ProperPart`). Such distribution characteristics can be used for steering qualitative diversification (thereby promoting CNGs to Bayesian networks), in order to generate synthetic test data sets exhibiting near real-world characteristics.

Qualitative diversification should be aware of error models. In our approach, CNGs steer the qualitative diversification of synthetic duplicates. Although such CNGs can be defined domain-independently, fitting them to the errors encountered in a particular domain is possible. Depending on real-world system factors, such as the type of user interface, various errors may occur: for example, values may be simply outdated (differ from the real value by some offset), or be entered with transposed digits. In road traffic management, for example, traffic jams are either detected by sensors, which may fail arbitrarily (e. g., a large traffic jam may be detected as two smaller, disrelated ones), or entered by humans, which may enter wrong data. Such errors should be represented by adapting the respective CNGs and adding or removing edges, so that errors are in the correct similarity neighborhood.

Characteristics of value ranges bound quantitative diversification. Value ranges provide a model of the world, and may be instantiated to represent different real-world spaces. For example, consider our value range for objects on highways being defined as intervals on a strictly monotonic, linear space. Each concrete highway is an instance of such a linear space that may differ in length from other instances. For quantitative diversification, we may use these additional characteristics as constraints, or deliberately ignore them to also create inconsistent synthetic duplicates.

Generalizing qualitative and quantitative diversification to other domains. Numerous causes for contradictory values in duplicates are known in other domains [16], for instance errors in strings and numbers such as typographical errors and synonyms. Since a representation of these causes as a CNG is possible, extending SemGen towards domains that rely on string representations will begin by defining an appropriate CNG together with relation derivation functions and their inverse functions.

References

1. J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.
2. N. Baumgartner, W. Gottesheim, S. Mitsch, W. Retschitzegger, and W. Schwinger. BeAware!—situation awareness, the ontology-driven way. *International Journal of Data and Knowledge Engineering*, 69(11):1181–1193, November 2010.
3. N. Baumgartner, W. Gottesheim, S. Mitsch, W. Retschitzegger, and W. Schwinger. Towards duplicate detection for situation awareness based on spatio-temporal relations. In *Proceedings of the 9th International Conference on Ontologies, DataBases and Applications of Semantics*, Crete, Greece, October 2010.
4. P. Bertolazzi, L. D. Santisy, and M. Scannapieco. Automatic record matching in cooperative information systems. In *Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03)*, 2003.
5. T. Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002.
6. N. Bruno and S. Chaudhuri. Flexible database generators. In *Proceedings of the 31st international conference on Very large data bases*, pages 1097–1107, 2005.
7. H. T. Bruns and M. J. Egenhofer. Similarity of spatial scenes. In M.-J. Kraak and M. Molenaar, editors, *Proceedings of the 7th International Symposium on Spatial Data Handling (SDH)*, pages 31–42, Delft, The Netherlands, August 1996.
8. D. Chays, S. Dan, P. G. Frankl, F. I. Vokolos, and E. J. Weber. A framework for testing database applications. *SIGSOFT Software Engineering Notes*, 25:147–157, August 2000.
9. F. Dylla and J. O. Wallgrün. On generalizing orientation information in OPRA$_m$. In *Proceedings of the 29th Annual German Conference on AI (KI2006)*, LNCS, pages 274–288, Bremen, Germany, August 2007.
10. C. Freksa. Conceptual neighborhood and its role in temporal and spatial reasoning. In *Proceedings of the IMACS International Workshop on Decision Support Systems and Qualitative Reasoning*, pages 181–187, Toulouse, France, March 1991.
11. C. Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1):199–227, 1992.
12. M. A. Hernández and S. J. Stolfo. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, SIGMOD Rec., pages 127–138, New York, NY, USA, 1995.
13. J. E. Hoag and C. W. Thompson. A parallel general-purpose synthetic data generator. *SIGMOD Rec.*, 36:19–24, March 2007.

14. K. Houkjær, K. Torp, and R. Wind. Simple and realistic data generation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 1243–1246, 2006.
15. W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7:81–99, 2003.
16. F. Naumann and M. Herschel. *An Introduction to Duplicate Detection*. Morgan & Claypool, 2010.
17. Randell, D.A., Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, October 1992.
18. M. Weis, F. Naumann, and F. Brosy. A duplicate detection benchmark for xml (and relational) data. In *SIGMOD 2006 Workshop on Information Quality for Information Systems (IQIS)*, Chicago, IL, USA, June 2006.